Eren Orbey  
May 2, 2019  
CPSC 490  
Dragomir Radev

*Toward a Text-Based Classification of USPTO Data*:  
Analyzing the Readability of Patent Grants from 2006 to 2015

**Abstract**

The digital archive of the United States Patent and Trademark Office (USPTO) provides a wealth of textual data that can be analyzed to classify patent documents according to their reading complexity. I calculate standardized readability scores for the plain-text elements of patent documents to determine how the clarity and concision of patent writing vary across time and industry. I introduce the question of whether readability predicts the speed at which filed patents are approved and published—a crucial metric for competitive inventors racing to lay claim to intellectual property. Patents assigned to International Patent Classification (IPC) Classes C (Chemistry; Metallurgy), G (Physics), and H (Electricity) rank consistently as the most difficult, least readable texts according to four widely used readability metrics, and correspond to longer delays between application and publication. The data also suggest a decline in patent readability between 2006 and 2015, as demonstrated by increasing mean Flesch indices and decreasing mean Flesch-Kincaid, SMOG, and Dale-Chall grade levels. Within each of the broadest international classification levels, further experimentation ought to be performed to determine a relationship between readability scores and time required for approval and publication.

**Introduction and Problem Description**

Technological innovations in the last decade have contributed to an increasing digitization of human interaction and communication in the form of text. Researchers, in turn, have introduced text as an input to data analysis across economic sectors. Shiller considered the so-called “epidemiology of narratives relevant to economic fluctuations” by relying on quantitative analysis (Shiller, 2017). Kelly employed textual analysis of high-dimensional data from patent documents to “create new indicators of technological innovation,” calculating quantitative indices to identify such breakthroughs and to predict productivity at the aggregate, sectoral, and firm levels (Kelly et. al., 2018). The archive of patent documents contains a wealth of data, though an underexplored quality of such data is the reading complexity of the grant text itself.

For decades, governments, businesses, and educators across the world have relied on so-called readability scores as objective proxies for the complexity of English-language textual documents. These formulas have provided minimum thresholds of clarity and comprehensibility for financial disclosure forms, standardized test questions, technical manuals, and medical inserts (Chall, 1977; Hengel, 2017). Higher readability scores correlate in some sectors with increased
readership (Richardson, 1977; Hengel, 2017). More readable articles in academic journals tend to be cited more frequently and to win more awards (Dowling, 2018; Sawyer, 2008).

This report seeks to apply common readability metrics to the text of archives from the USPTO to determine whether and how the clarity and concision of patents vary across industry and time. Among published patents, are more readable documents likely to be approved more quickly? Might style matter as much as substance in the sphere of intellectual property?

Approach

I use various text-based readability metrics to assess differences in clarity and concision in the abstracts of USPTO patent grants from 2006 to 2015. In order to standardize this analysis, I adhere to the hierarchical IPC system, grouping grants according to their industry designations, which may be liable to increase or decrease the overall readability of an entire classification of patents. (For example, due to industry-specific vocabulary, patents classified under “Human Necessities” may well be uniformly more readable than those classified under “Electricity.”) I examine these calculated readability scores to assess how the clarity and concision of patent writing vary across time and industry. I also consider the time elapsed between the filing and the publication of each patent to determine whether and how the clarity and concision of grants affect their passage toward approval by the USPTO bureaucracy. The purpose of this approach is to investigate the efficacy of classifying patents according to the comprehensibility of their grant text. Is there an incentive for inventors to file clearer, less complex patent grants—even though clarity might decrease the barrier to entry for future competitors?

Data Used

The data include the full text of each grant issued weekly from January of 2006 to March of 2015, courtesy of an agreement between Google and the USPTO that makes bulk downloads of patent data available at no charge. Every Tuesday, a new bulk file is released, usually containing between two and five thousand patents granted on the same day. (Bulk files are not updated once published.) Each file consists of a concatenation of the Standard Generalized Markup Language (SGML) in accordance with the U.S. Patent Grant Version 2.4 Document Type Definition (DTD) and eXtensible Markup Language (XML) in accordance with the U.S. Patent Grant Version 2.5 DTD. A number of existing libraries facilitate USPTO data access:

- **Patent Parsing Tools** provides tools for generating training and test set from Google’s USPTO data helpful with for testing machine learning algorithms
- **PatentsView API** can take in a list of values (such as patent numbers), retrieve multiple data points, and then convert and merge the results into a CSV file.
Most of the data is available for free manual download here. I have downloaded the entire available archive of XML releases from 2006 to 2015 to the following directories:

- /home/lily/eo235/data (tangra.cs.yale.edu)
- /home/accts/eo235/cs490 (node.zoo.cs.yale.edu)

**Evaluation Method**

I. *Parsing Patent Data*

Several existing libraries purport to parse the USPTO releases, though many of them either are outdated themselves or rely on outdated document type definitions. To avoid discrepancies, I analyze a span of time, from 2006 to 2015, in which the USPTO data conforms to an identical schema, and I rely on my own Python script to reformat and parse each concatenated XML release. My script fixes the concatenated nodes of each release to a root element and uses the ElementTree API to extract (or, in one case, calculate from extracted values) the following fields for each patent:

- **invention_title** (string)
  - e.g. “Sanitary finger cap”
- **application_date** (string converted to datetime)
  - e.g. June 6, 2009
- **publication_date** (string converted to datetime)
  - e.g. February 3, 2015
- **delta_in_days** (int calculating the time between application and publication)
  - e.g. 2062
- **abstract** (string cleaned as plain text—see II. Cleaning Text of Abstracts)
  - e.g. “A sanitary finger cap including: an insertion portion having a generally cylindrical shape and adapted to be fitted to a third knuckle of a finger connected to the center of a hand or to a portion of a second knuckle of the finger; and a grasping portion connected with one side of the insertion portion as a unitary body, in a shape of a general corn or pyramid in such a manner as to become small in thickness as it goes toward the other side thereof, and adapted to be fitted to a first knuckle as the end portion of the finger or to the second knuckle of the finger so as to grasp food thereon.”
- **ipcr_classification** (string consisting of a single character A-H, as below)
  - A HUMAN NECESSITIES
  - B PERFORMING OPERATIONS; TRANSPORTING
  - C CHEMISTRY; METALLURGY
  - D TEXTILES; PAPER
E  FIXED CONSTRUCTIONS
F  MECH. ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
G  PHYSICS
H  ELECTRICITY
  o  e.g. “A”

II.  Cleaning Extracted Abstract Text

My script extracts the abstract from each patent XML element by using an iterator to return all inner text within the appropriate node. (This text tends to be just one paragraph, but, because that paragraph may be nested within additional tags, the use of an iterator assures no other plain text classified under a patent’s abstract field goes unparsed.) The text is stripped of new lines and then subjected to further cleaning by the Textatistic and PyHyphen libraries, which perform the following operations:

- Replacing em and en dashes with hyphens
- Removing hyphens in hyphenated single words (e.g. co-author)
- Removing decimals
- Removing mid-sentence rhetorical punctuation (e.g. “[...] (must polluters pay?).”)
- Replacing common abbreviations with their full text

Such modifications to the text, though they are not exhaustive, aim toward a level of standardization without which periods might skew sentence counts (among other potential malfunctions).

III.  Assigning Readability Scores

Advanced vocabulary, sentence length, and grammatical complexity are dependable predictors of text difficulty (Hengel, 2017). Though hundreds of formulas rely on these relationships to measure readability, I focus on four of the most widely used and tested metrics for adult reading material: Flesch Reading Ease, Flesch-Kincaid, Simple Measure of Gobbledygook (SMOG), and Dale-Chall. The scores—which can be calculated manually or, more simply, by libraries like Textatistic and textstat—rely on the following counts from each text sample.

- `char_count` (int)
  o  The number of non-space characters.
- `sybl_count` (int)
  o  The overall number of syllables.
- `word_count` (int)
  o  The overall number of words.
• **polysyllblword_count** (int)
  o The number of words containing three or more syllables.

• **notdalechall_count** (int)
  o The number of words not found on a list of 3,000 words understood by eighty percent of fourth-grade readers, according to the 1995 findings of Chall and Dale.

• **sent_count** (int)
  o The number of sentences.

The readability indices themselves are described below. The formula to which each corresponds appears in the following table.

• **Flesch Reading Ease**: This index is perhaps the most widely used measure. It ranks passages, most of which score between 0 and 100, in ascending order of readability. More readable passages yield higher scores; more difficult passages yield lower scores. *(Reader’s Digest tends to rank in the 60s; the Harvard Law Review in the 30s.)* The index originated under contract for the U.S. Navy and was first used to assess the difficulty of military manuals in the 1970s. Many states mandate that insurance policies and other common legal paperwork reach a certain threshold of readability.

• **Flesch-Kincaid**: This index, developed alongside the former, estimates the number of years of formal education required to understand an evaluated text—so, in this case, more readable passages earn lower scores. (Numerically, there is no upper bound.)

• **SMOG**: This index also estimates the number of years of formal education required to understand an evaluated text. Its assumption that polysyllabic words are categorically more difficult is perhaps one of the reasons it is less commonly used.

• **Dale-Chall**: This index, too, estimates the number of years of formal education required to understand an evaluated text. It classifies difficult words as those that do not appear on a list the researchers determined to be familiar to most fourth graders. The original list, published in 1948, when the metric was devised, included about 800 familiar words. A revised list includes 3,000.

**FIGURE 1**: Readability Scores and Corresponding Formulas. Recall that whereas higher Flesch scores suggest greater ease of readability, higher Flesch-Kincaid, SMOG, and Dale-Chall grade levels suggest the inverse: *more* years of education required to read an evaluated text.
**Results**

This run of the experiment parsed and processed a total of 553,199 patents. I use the [pandas](https://pandas.pydata.org) and [Matplotlib](https://matplotlib.org) libraries to present and analyze data from the outputted CSVs of parsed data.

**FIGURE 3**: Patents Processed by Year. The script examined USPTO releases from each month of Google’s bulk downloads between January of 2006 and March of 2015.

<table>
<thead>
<tr>
<th>Year</th>
<th>Patents Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>46598</td>
</tr>
<tr>
<td>2007</td>
<td>38284</td>
</tr>
<tr>
<td>2008</td>
<td>42262</td>
</tr>
<tr>
<td>2009</td>
<td>39357</td>
</tr>
<tr>
<td>2010</td>
<td>55663</td>
</tr>
<tr>
<td>2011</td>
<td>59467</td>
</tr>
<tr>
<td>2012</td>
<td>64782</td>
</tr>
<tr>
<td>2013</td>
<td>71266</td>
</tr>
<tr>
<td>2014</td>
<td>76021</td>
</tr>
<tr>
<td>2015</td>
<td>59419</td>
</tr>
</tbody>
</table>
**FIGURE 4**: Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) and Mean Time to Publication (days) by Year (2006 to 2015). The decreasing Flesch indices and increasing Flesch-Kincaid, SMOG, and Dale-Chall grade levels suggest a general decline in readability (i.e. increase in textual difficulty and complexity).

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean Readability Scores</th>
<th>Mean Time to Publication (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flesch</td>
<td>Flesch-Kincaid</td>
</tr>
<tr>
<td>2006</td>
<td>32.05</td>
<td>18.50</td>
</tr>
<tr>
<td>2007</td>
<td>31.26</td>
<td>18.72</td>
</tr>
<tr>
<td>2008</td>
<td>30.31</td>
<td>18.95</td>
</tr>
<tr>
<td>2010</td>
<td>29.39</td>
<td>19.28</td>
</tr>
<tr>
<td>2011</td>
<td>28.22</td>
<td>19.67</td>
</tr>
<tr>
<td>2012</td>
<td>27.72</td>
<td>19.81</td>
</tr>
<tr>
<td>2013</td>
<td>27.08</td>
<td>20.00</td>
</tr>
<tr>
<td>2014</td>
<td>26.84</td>
<td>20.08</td>
</tr>
<tr>
<td>2015</td>
<td>26.68</td>
<td>20.17</td>
</tr>
</tbody>
</table>

**FIGURE 5**: Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) by Year (2006 to 2015). Recall that more readable texts produce higher Flesch scores and lower Flesch-Kincaid, SMOG, and Dale-Chall grade levels—that is, the small decrease in mean Flesch scores and smaller increase in the others demonstrates a slight decline in patent readability. The texts seem to become more complex.
FIGURE 6: Mean Time to Publication (days) by Year (2006 to 2015). The data suggest a lag in time for publication toward the middle of the analyzed decade.

FIGURE 7: Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) and Mean Time to Publication (days) by IPC Class (A-H). The readability scores rank Classes C (Chemistry; Metallurgy), G (Physics), and H (Electricity) as consistently more difficult texts. The correlation between readability and time to publication is not definitive across IPC classes—consider that Class A (Human Necessities) shows a high time to publication despite being relatively readable on average. It is notable, though, that Class C, statistically the least readable, shows the high mean time to publication.

<table>
<thead>
<tr>
<th>IPC Class (A–H)</th>
<th>Mean Readability Scores</th>
<th>Mean Time to Publication (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flesch</td>
<td>Flesch Kincaid</td>
</tr>
<tr>
<td>A</td>
<td>32.83</td>
<td>17.72</td>
</tr>
<tr>
<td>B</td>
<td>31.94</td>
<td>19.69</td>
</tr>
<tr>
<td>C</td>
<td>24.62</td>
<td>20.04</td>
</tr>
<tr>
<td>D</td>
<td>31.34</td>
<td>19.86</td>
</tr>
<tr>
<td>E</td>
<td>39.08</td>
<td>17.30</td>
</tr>
<tr>
<td>F</td>
<td>34.43</td>
<td>18.88</td>
</tr>
<tr>
<td>G</td>
<td>26.11</td>
<td>19.99</td>
</tr>
<tr>
<td>H</td>
<td>26.94</td>
<td>20.02</td>
</tr>
</tbody>
</table>
FIGURE 8: Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) by IPC Class (A-H). Classes C (Chemistry; Metallurgy), G (Physics), and H (Electricity) rank consistently the most difficult according to lower Flesch scores (blue bars) and higher Flesch-Kincaid, SMOG, and Dale-Chall grade levels (green, yellow, and red bars).

FIGURE 9: Mean Time to Publication (days) by IPC Class (A-H). Note that Class C (Chemistry; Metallurgy), the least readable according to the prior figures, requires the greatest mean time to publication.
FIGURE 10: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class A (Human Necessities).

FIGURE 11: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class B (Performing Operations; Transporting).
FIGURE 11: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class C (Chemistry; Metallurgy).

FIGURE 12: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class D (Textiles; Paper).
FIGURE 13: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class E (Fixed Constructions).

FIGURE 14: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class F (Mech. Engineering; Lighting; Heating; Weapons; Blasting).
FIGURE 15: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class G (Physics).

FIGURE 16: Time to Publication (days) by Flesch Scores (blue), Flesch-Kincaid Grade Levels (green), Smog Grade Levels (yellow), and Dale-Chall Grade Levels (red) for IPC Class H (Electricity).
In Figures 10 through 16, the scatter plots corresponding to each IPC Class (A-H) exhibit variance that does not allow for conclusive regression analysis between each readability score and time to publication. It is notable, though, that the graphs for each individual readability score across classes (i.e. the same-colored scatter plots) conform to similar distributions.

**Discussion and Conclusion**

The data suggest a decline in patent readability between 2006 and 2015, as demonstrated by increasing mean Flesch indices and decreasing mean Flesch-Kincaid, SMOG, and Dale-Chall grade levels (Figures 4 and 5). Patents assigned to IPC Classes C (Chemistry; Metallurgy), G (Physics), and H (Electricity) rank consistently as the most difficult, least readable texts according to all four metrics and correspond to longer delays between application and publication (Figures 7, 8, and 9). Within each IPC Class (A-H), the data does not conclusively indicate a direct or inverse variation between calculated readability scores and time required for publication.

Further experimentation should be performed to confirm observed relationships and to attain more results with less variance over a longer time span. The existing script can be modified to address the following open areas of exploration:

- Though the abstract section is the most consistent plain-text field in length and in tone across patents from different IPC Classes, it is also occasionally shorter than the “claims” section and usually shorter than the “description” section. A future run of the extraction script could calculate and plot the same readability scores on each of these sections as well.

- The plain-text sections of many patent grants feature images that enhance or clarify descriptions. A future run of the extraction script could exclude all such patents before calculating readability scores; alternately, it could compare patents that feature images with those that do not to determine whether the presence of images compensates for the quality of less readable documents.

- The IPC Class schema contains subfields and subclasses that further differentiate patent grants beyond the high-level A through H distinctions. Recreating Figures 10 through 16 with more specific classifications may evince a more conclusive relationship between calculated readability and time required for publication.

- The USPTO full-text archives date back to 1967, so, though the XML tags and classification schemas vary widely over time, a future experiment could expand the described approach and the existing script to assess the full range of grant data, provided that it manages to classify patents from different decades as consistently as possible.
References


Appendix: Code and Data

The available archive of XML releases from 2006 to 2015 can be found in the following directories, where CSV spreadsheets with parsed and extracted data, as well as sample extraction and analysis scripts, also appear.
• /home/lily/eo235/data (tangra.cs.yale.edu)
• /home/accts/eo235/cs490 (node.zoo.cs.yale.edu)